Generative Large Language Models in Automated Fact-Checking: A Survey

Anonymous ACL submission

Abstract

The dissemination of false information on online platforms presents a serious societal challenge. While manual fact-checking remains crucial, Large Language Models (LLMs) offer promising opportunities to support fact-checkers with their vast knowledge and advanced reasoning capabilities. This survey explores the application of generative LLMs in fact-checking, highlighting various approaches and techniques for prompting or fine-tuning these models. By providing an overview of existing methods and their limitations, the survey aims to enhance the understanding of how LLMs can be used in fact-checking and to facilitate further progress in their integration into the fact-checking process.

1 Introduction

The modern digital era has introduced numerous challenges, including the widespread dissemination of false information, a problem exacerbated by the rise of social media. Fact-checking is a key strategy for combating false information (Vlachos and Riedel, 2014), but it is still largely carried out manually by fact-checkers. As the number of fact-checkers remains insufficient to keep pace with the growing volume of misinformation (Aïmeur et al., 2023), efforts are underway to develop automated fact-checking systems. These systems leverage artificial intelligence and LLMs to reduce the work of fact-checkers (Nakov et al., 2021a).

The application of LLMs in automated fact-checking has demonstrated significant potential for improving both the efficiency and accuracy of fact-checking (Wang et al., 2023a). LLMs are trained on large-scale datasets, and they leverage billions of parameters to capture nuances and patterns of natural language. Moreover, generative LLMs, specifically designed for text generation, open new possibilities for integrating LLMs into the verification process by generating meaningful, natural text that was not possible before and automating complex reasoning tasks.

Previous surveys on fact-checking in NLP domain have primarily focused on the needs of fact-checkers (Nakov et al., 2021a; Hrckova et al., 2024), specific fact-checking tasks (Guo et al., 2022b), datasets (Guo et al., 2022a) or traditional and BERT-based methods (Thorne and Vlachos, 2018; Zeng et al., 2021). One recent survey addresses combating misinformation using LLMs, illustrating both the opportunities and challenges they present (Chen and Shu, 2023b). However, this study lacks a detailed exploration of the methods employed. This presents a gap for a more thorough examination of generative LLMs' role in fact-checking.

Our main contribution is a comprehensive overview of approaches and limitations in using generative LLMs for automated fact-checking. We review 69 papers, highlighting relevant methodologies and innovative prompting techniques for researchers exploring LLM-aided information verification. We analyze fact-checking tasks, LLM methods, common techniques and languages covered in surveyed papers. Additionally, we outline future challenges and potential directions for advancing the use of generative LLMs in the verification process.

2 Methodology

2.1 Selection

To compile a set of relevant articles, we conducted a manual search of papers presented at *ACL*, *EACL*, *NAACL* and *EMNLP* conferences. Additionally, we performed a keyword search using Google Scholar and ArXiv. We expanded this collection by incorporating papers cited in related works and those that cited any article from our initial list. Furthermore, we included papers from the *CheckThat! Shared tasks*. In total, we gathered 69 papers that employ

generative LLMs for fact-checking. Additional details about our selection process are provided in Appendix A.

2.2 Observed Aspects

In our survey, we categorized the papers based on several key aspects. First, we investigated fact-checking tasks addressed in each study (Section 3) and the methods employed to incorporate LLMs for tackling these tasks (Section 4). These methods represent a high-level abstraction of the approaches used, categorized based on the output type generated by LLMs. Another aspect we examined is techniques (Section 5), which refer to the strategies, such as prompting and fine-tuning, used to adapt and apply LLMs for specific fact-checking tasks to ensure they generate desired outputs. Given that false information is a global issue, we also analyzed languages covered in each study (Section 7). Furthermore, we classified what type of knowledge base was used for *evidence retrieval* (see Table 5) and whether LLMs utilized evidence for the *fact-verification* task (see Table 6). Relevant studies for each task are summarized in Tables 3 to 6.

2.3 Scope of the Survey

This survey is intended for researchers with expertise in NLP and a fundamental understanding of methods across the field. We aim to present an overview of techniques employed in fact-checking using generative LLMs. While fact-checking has evolved to cover multiple modalities, our study is confined to textual and tabular data analysis. Unlike surveys that provide detailed descriptions of fact-checking datasets, our focus is on the methods, techniques, and challenges of leveraging generative LLMs. Our aim is to focus on the verification of the truthfulness of a given input using LLMs, rather than on evaluating LLM outputs for hallucinations (Ji et al., 2023) or factuality (Wang et al., 2023a, 2024).

3 Fact-Checking Tasks

Fact-checking is the process of verifying the truthfulness of a given claim (Guo et al., 2022b). This process involves several distinct tasks, each essential for determining the credibility of the information. In this section, we sort identified tasks based on the number of related papers and highlight the most common methods used for each task.

3.1 Fact Verification & Fake News Detection

Fact verification and *fake news detection* are the most prominent tasks when incorporating LLMs into the fact-checking process, with 53 out of 69 surveyed papers aimed at these tasks. *Fact verification* involves assessing the veracity of a given claim or few sentences. In contrast, *fake news detection* aims to check the trustworthiness of longer texts, such as news articles (Li and Zhou, 2020).

A key approach (44 out of 53 papers) in *fact verification* is to use LLMs to classify claims into several categories, which can range from two (true or false) to three (supports, refutes, or not enough information) or even more classes (Buchholz, 2023; Hoes et al., 2023a). In addition to classification, **explanation generation** (22×) is commonly employed in conjunction with generative LLMs (Althabiti et al., 2023; Russo et al., 2023b; Cekinel and Karagoz, 2024; Kao and Yen, 2024).

3.2 Evidence Retrieval

Thirteen research papers have addressed the task of *evidence retrieval* using LLMs. *Evidence retrieval* is about gathering essential information from trusted sources to assess the veracity of claims. The evidence can take various forms, such as text or tables. This evidence serves as the knowledge necessary for accurate fact-checking. The selection of an appropriate knowledge source is crucial, as Wikipedia and the open web exhibit several weaknesses. In contrast, scientific reports (Bhatia et al., 2021) offer a more credible alternative.

The predominant approach for *evidence retrieval* using LLMs is ranking (5 out of 13) (Jiang et al., 2021; Pradeep et al., 2021, 2020). Additionally, alternative strategies have explored the capabilities of LLMs to **generate queries** (1×) for web retrieval (Prieto-Chavana et al., 2023), and **rationale selection** (5×) (Wang et al., 2023b; Evans et al., 2023; Kamoi et al., 2023b), in which LLMs are tasked with identifying relevant sentences from the retrieved documents.

3.3 Claim Detection

Eleven of 69 papers addressed the *claim detection* task, indicating a moderate emphasis. *Claim detection* concerns identifying claims containing verifiable information necessitating verification. These claims, often termed **check-worthy** or **verifiable** claims, can include statements that may be false, misleading, or potentially harmful to society.

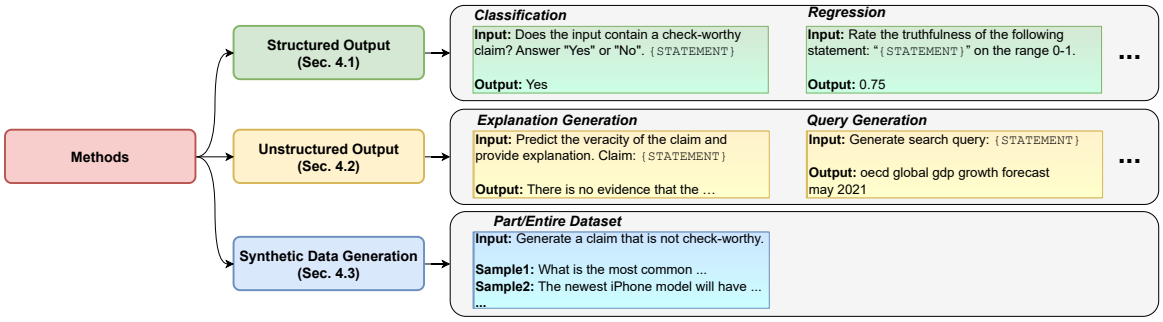

Figure 1: A taxonomy of methods for integrating generative LLMs in fact-checking, illustrating examples of model inputs and outputs. The methods employed with generative LLMs are classified based on the output type.

Typically, *claim detection* is approached as a binary classification (7 out of 11), as LLMs are primarily employed as binary classifiers. Additionally, several researchers have explored the generative capabilities of LLMs to **generate claims** (4×) (Chen et al., 2022; Gangi Reddy et al., 2022), allowing them to produce a set of verifiable claims based on a given text, or **generate questions** (1×) (Chen et al., 2022), aiming to formulate queries for verification. Leveraging these generative techniques allows LLMs to extract key information and decompose complex statements into easily verifiable components or questions.

3.4 Previously Fact-Checked Claims Detection

The task of *previously-checked claims detection* aims to reduce redundant work for fact-checkers. Currently, it is the least explored area using LLMs, with only five identified papers. However, it is essential for fact-checkers to ascertain whether a claim has been previously fact-checked, as misleading claims propagate across various sources and languages (Pikuliak et al., 2023b). Thus, the goal of *previously fact-checked claims detection* is to identify similar claims from a database that have already undergone fact-checking.

In the context of *previously fact-checked claims detection*, LLMs are commonly employed to rank or filter retrieved fact-checks using the information retriever systems (3×) (Shaar et al., 2020; Shliselberg and Dori-Hacohen, 2022; Neumann et al., 2023). Another approach is to use LLMs to classify the pairwise relationships between claims and retrieved fact-checks as textual entailment (2×) Choi and Ferrara (2023, 2024).

4 Taxonomy of Methods

In this section, we discuss the methods of using LLMs to perform various fact-checking tasks. **We categorize methods based on their *output type*.** This taxonomy is illustrated in Figure 1: (1) *Structured output* is used when LLMs process samples and generate answers in a predefined structure: a class for a classification task, a number for a regression task, ranking of inputs, etc. In this case, LLMs often replace older specialized approaches. (2) *Unstructured output* is used when LLMs process samples and generate texts such as explanations, summaries, etc. that are used in the fact-checking process. (3) *Synthetic data generation* is an approach when LLMs generate new samples for our datasets. These datasets can then be used to train other (often smaller and specialized) models.

4.1 Structured Output

Methods that produce structured output are the most common in fact-checking, appearing in 59 out of 69 papers. These methods are advantageous because they allow easier evaluation, offer higher execution efficiency and easier fine-tuning due to the availability of datasets with labeled targets. This makes them particularly beneficial for fact-checking tasks. In this section, we further categorize these methods into three output types: classification, regression and ranking.

Classification. LLM-based classification utilizes LLMs to predict categorical labels from a predefined set, e.g., veracity labels. In this approach, LLMs are tasked to select the most probable class. Classification is commonly employed in fact-checking (54 out of 69), as **many tasks involve categorizing claims**, especially in *claim detection* (yes/no) or *fact verification* (e.g., false, half true,

true). It is also effective in *detecting previously fact-checked claims*, as an entailment classification between a given statement and previously fact-checked claims (Choi and Ferrara, 2023, 2024).

Regression. In regression (7×), LLMs generate predicted scores across various scales (e.g., 0-1, 0-100). However, in fact-checking, **regression is often redefined as a classification problem** due to the absence of datasets with score targets (Li et al., 2023b; Vergho et al., 2024; Setty, 2024). In this scenario, LLMs predict a score from the selected range, which is converted into a classification based on a given threshold, typically 50% for binary classification. This approach is beneficial for its ability to optimize thresholds based on identified LLM characteristics (Pelrine et al., 2023). Regression is widely used in *previously fact-checked claims detection* and *evidence retrieval*, where claims or pieces of evidence are ranked by predicted scores. It is also applied, although less frequently, in *fact verification* (Guan et al., 2023; Jiang et al., 2023; Evans et al., 2023), where the goal is to predict the truthfulness score of a claim.

Ranking. Ranking (8×) involves sorting claims, documents, or pieces of evidence based on their relevance to a given query. While less common in fact-checking than other methods, it can still play a significant role. One approach we define as **classification ranking** uses the logits of particular tokens (e.g., `true`) produced by the LLM (Pradeep et al., 2021). In this case, the LLM predicts a particular class, and the probability associated with that class is then used for ranking.

Another approach involves **rationale selection** (5×), where LLMs are directly instructed to rank evidence from a prompt. In this method, LLMs identify and select the most relevant parts of the retrieved evidence or their own knowledge, such as by selecting relevant sentences (Evans et al., 2023; Tan et al., 2023) or relevant rationales (Wang et al., 2023b).

4.2 Unstructured Output

Methods with unstructured outputs typically involve generating continuous text, which poses greater challenges in assessment. However, these methods provide more detailed responses, often including justifications, which enhance fact-checking by offering insights into the LLM's reasoning. Since generative LLMs produce open-ended outputs rather than selecting from a set of labels, they can address a broader range of fact-checking tasks.

We categorized these methods based on the output type generated by LLMs into three categories: (1) coherent, meaningful text consisting of multiple sentences; (2) single-sentence outputs; and (3) outputs consisting of only a few keywords or phrases.

Explanation Generation. Explanation generation (22×) is a method of generating textual explanations that clarify the reasoning behind an LLM's output, particularly why a specific claim was categorized in a certain way, such as being classified as false (Russo et al., 2023c). The generated output is commonly meaningful, continuous text that offers insights into the model's decision-making process.

In addition to directly prompting LLMs to explain their decision, **summarization can also be used as a form of explanation generation**, in tasks like *fact-verification* (Russo et al., 2023b). However, the limitation of summarization lies in its tendency to omit crucial details from the retrieved evidence, leading to summaries that may not fully address misleading claims, nor be as persuasive as traditional fact-checking articles.

Claim and Question Generation. The goal of claim generation is usually to produce one sentence that captures all key information for various tasks, including *claim detection*, *evidence retrieval* or *fact-verification*. The claim generation can also be known as **claim normalization** (Sundriyal et al., 2023), where the aim is to extract a single normalized claim from a social media post. However, challenges arise when multiple claims are present in the input, complicating the creation of a single normalized claim. Therefore, claim generation can also produce a list of claims.

A part of claim generation is **span detection** (Gangi Reddy et al., 2022), which aims to identify the exact boundaries of the claim within the text. The advantage of span detection lies in its ability to select the precise wording of the claim.

Claim generation and **question generation** share a similar output format – a single sentence. However, in question generation, the output specifically consists of a question, mostly used to verify a given claim (Chen et al., 2022).

Query Generation. Another form of unstructured output is query generation (Prieto-Chavana et al., 2023), where the output comprises several words or short phrases primarily used for retrieving

evidence from external sources, such as search engines. This method allows LLMs to extract essential parts of a claim, enhancing retrieval efficiency and minimizing irrelevant results that may arise from using the full claim.

Next hop prediction (Malon, 2021) is a specific type of query generation aimed at identifying and generating the title with a sentence of the evidence to be retrieved. This is especially important when previously retrieved evidence introduces new entities that require further elaboration.

4.3 Synthetic Data Generation

In synthetic data generation, LLMs are employed **to create entire datasets or their parts**. Unlike the generation from Section 4.2, where LLMs are used to generate results, synthetic data generation creates data that are then used for training or fine-tuning. This method proves valuable when datasets are either unavailable or limited for specific tasks or languages. In fact-checking, it can facilitate the creation of datasets containing disinformation (Chen and Shu, 2023a; Vykopal et al., 2024) or fake-news (Jiang et al., 2023; Huang and Sun, 2023).

Synthetic data generation can **augment datasets by filling gaps in underrepresented categories or increasing variability**. For instance, LLM can generate debunked claims to extend datasets (Singh et al., 2023), produce emotional misinformation tweets (Russo et al., 2023a) or create counterparts for claim-evidence pairs with contrasting veracity and high word overlap (Zhang et al., 2024a).

5 Techniques

In this section, we analyze how different studies applied techniques to guide LLMs in achieving accurate outputs. We observed several popular techniques to improve the performance of LLM, which we categorized into three groups: ***prompting***, ***fine-tuning*** and ***augmentation with external knowledge***. These techniques can overlap, with some studies using combinations, such as linking prompts with external resources. Papers were classified by the most prominent technique; for example, if prompting was combined with external knowledge, they were categorized under augmentation with external knowledge.

5.1 Prompting

Prompting is a set of techniques used to improve the performance of LLMs by designing the instructions. **This approach is versatile and particularly effective in scenarios with limited data**, such as fact-checking, where datasets are often scarce and labeled data is difficult to obtain.

However, prompting faces several limitations, especially with complex tasks like in fact-checking. For instance, *detecting previously fact-checked claims* involves ranking, where prompting alone may struggle to produce accurate results (Choi and Ferrara, 2024). Another limitation is the necessity for clear instructions and contextual descriptions to ensure accurate task execution. **To reduce ambiguities, it is essential to define all relevant terminology within the prompt**, e.g., the attributes of *check-worthy claims*. Therefore, prompt engineering is important, and LLMs can assist in prompt creation (Cao et al., 2023).

Prompt structure and wording can influence LLM performance. Techniques like role specification (Li and Zhai, 2023) and JSON formatting (Ma et al., 2023) have been shown to enhance performance. Role specification assigns the LLM a defined role (e.g., fact-checker), improving task comprehension, while JSON formatting structures instructions and outputs in a consistent manner (e.g., `Check the claim: {claim}\nAnswer:`). Additionally, **output sequencing affect performance** (Vergho et al., 2024; Pelrine et al., 2023). Presenting the explanation first encourages LLM reasoning, leading to more accurate assessments, whereas placing the final verdict first may cause the LLM to focus on justifying the outcome rather than producing a truthful explanation.

Fact-checking is challenging for prompting LLMs, as their responses can be inaccurate. Zeng and Gao (2023) addressed this by using multiple prompt variants, differing by a single word, to represent different relationships to the claim (e.g., true, unclear, or false). This method helps ensure consistency in the LLM's output, as small changes can affect predictions. Moreover, **LLMs can be prompted to rewrite claims in a neutral style**, improving evaluation reliability (Wu and Hooi, 2023) and resilience to adversarial attacks.

Prieto-Chavana et al. (2023) used **prefixes** like `Search query:` **to guide LLMs in generating search queries** for *evidence retrieval*. Furthermore, by combining query generation with rationale selection, relevant documents are retrieved, and then the most pertinent information is selected.

Few-Shot Prompting. Few-shot prompting (23×) is a technique, where the model is provided

with multiple in-context examples, enhancing the LLM's understanding of the task. This approach can mitigate the limitations of standard prompting, where LLMs may struggle to fully comprehend the task based only on the given instructions. **Strong performance, however, depends on the quality of the examples employed**. During inference, LLMs can effectively learn the nuances of a task from labelled examples (Brown et al., 2020). Previous studies have also explored how the number of examples affects LLM performance (Cao et al., 2023; Zeng and Gao, 2023, 2024).

In-context examples are particularly useful in methods such as span detection (Gangi Reddy et al., 2022) or claim generation (Kamoi et al., 2023b; Li et al., 2023a). **To reduce false positives, it is crucial that the examples capture a broad range of characteristics** typical for the given task. Decomposing input text into sub-claims helps break down complex statements into smaller, verifiable components. However, during this process, LLM may unintentionally reformulate claims, potentially altering their meaning or veracity labels. By providing varied demonstrations, few-shot prompting can aid LLMs in comprehending the variability and defining characteristics of check-worthy claims.

Few-shot prompting extends beyond *claim detection*. It is also employed in *evidence retrieval*, such as generating search queries (Li et al., 2023a) or selecting relevant rationales (Wang et al., 2023b).

Chain-of-Thought. The Chain-of-Thought (CoT) technique (Wei et al., 2024) **leverages LLMs' reasoning abilities by incorporating intermediate reasoning steps** before generating predictions. This technique can be employed in both zero-shot (10 ×) (Cao et al., 2023; Chen and Shu, 2023a) or few-shot (6×) (Sawiński et al., 2023; Choi and Ferrara, 2023; Pan et al., 2023a; Wang and Shu, 2023; Zhang and Gao, 2023; Liu et al., 2024a) settings. In few-shot CoT, authors present examples that illustrate the reasoning process involved, such as follow-up questions for *claim detection* (Sawiński et al., 2023) or examples for normalized claim (Sundriyal et al., 2023).

5.2 Fine-Tuning

Fine-tuning (33×) is a method of adapting a pre-trained model to a specific task or use case. This technique is especially beneficial when **models lack the necessary capabilities to effectively address a given task through prompting**, particularly in complex domains where LLMs may have limited expertise. While fine-tuning can enhance performance, it is typically a matter for smaller LLMs due to the number of parameters. In contrast, larger LLMs often possess enhanced capabilities for tackling a wide range of tasks without additional fine-tuning, making the process less feasible and often unnecessary.

Single-Task Fine-Tuning One approach to fine-tuning involves fine-tuning on a single task (Agrestia et al., 2022; Petroni et al., 2021; Bhatia et al., 2021; Hyben et al., 2023; Althabiti et al., 2023; Russo et al., 2023b; Zeng and Zubiaga, 2024), which **enhances the model's specialization and performance in that area**. However, this method may lead to the phenomenon of catastrophic forgetting, where LLMs lose the ability to perform other tasks, even those within the same domain (Luo et al., 2024). Consequently, **addressing multiple tasks requires fine-tuning a specific LLM for each task**. A strategy for single-task fine-tuning involves classification ranking (Pradeep et al., 2021). This approach has been extended in previous works, e.g., by including named entities within the claim (Jiang et al., 2021), reranking of previously retrieved documents (Pradeep et al., 2021) or predicting the reliability of the retrieved passage based on a given claim (Fernández-Pichel et al., 2022). Classification ranking can be employed to rank evidence in *evidence retrieval* and to extract pertinent sentences from retrieved evidence (Pradeep et al., 2020).

Multi-Task Fine-Tuning. An alternative approach to single-task fine-tuning involves employing multiple tasks during training (Du et al., 2022). **This approach aims to create more generalizable model by exploiting knowledge transfer across tasks**. While the primary goal is not to enhance performance for a single task, each task can positively affect the others. Additionally, this approach allows the LLM to be specialized for multiple tasks simultaneously.

Prompt Style for Fine-Tuning. LLMs can be fine-tuned using target labels and various prompts. For fine-tuning encoder-decoder LLMs, researchers commonly employ prompts formatted as `"Claim: c Evidence: e Target:"` (Sarrouti et al., 2021). In the case of decoder-only LLMs, the input often consists only of the claim (Sawiński et al., 2023). Given that contemporary LLMs

are frequently instruction-tuned, **incorporating instructions along with the statement is a standard practice**. Notably, decoder-only models generally outperform encoder-decoder models, especially when trained with instructional inputs.

Use-cases. In fact-checking, **fine-tuning is mostly utilized for ranking tasks**, e.g., *previously fact-checked claims detection* and *evidence retrieval*. It also extends to other applications, including generating sub-questions (Chen et al., 2022) and next-hop prediction (Malon, 2021).

5.3 Augmentation with External Knowledge

Since information changes over time and LLMs often lack up-to-date information, previous techniques are frequently combined with external tools and knowledge bases. This augmentation can involve using previously retrieved evidence, web engines, or external databases.

Evidence in Prompt. Many approaches integrate evidence from *evidence retrieval* into prompts to fill LLM knowledge gaps. We refer to the use of external knowledge as the **open-book setting** (Schlichtkrull et al., 2023), while relying only on internal knowledge is **closed-book settings**.

Given that **multiple pieces of evidence are often available to verify claim's veracity** (Jiang et al., 2021), two strategies are used: (1) verifying the claim sentence by sentence and aggregating the predicted labels for the final prediction; or (2) including all relevant evidence within a single query. The latter method tends to outperform the former, as **consolidating all necessary evidence into one prompt leads to more accurate predictions**. Beyond textual evidence, tabular data can also serve as a source, and it is typically linearized before being fed into LLMs (Zhang et al., 2024b).

Using External Tools. Besides incorporating retrieved evidence into prompts, **LLMs often interact with external tools**, such as search engines or knowledge bases (Zhang and Gao, 2023; Yao et al., 2023; Quelle and Bovet, 2023). This technique provides several advantages, most notably continuous access to up-to-date information, which is essential for accurate fact-checking. However, the effectiveness of these methods depends on the careful selection of credible sources.

Retrieval-Augmented Generation. An effective strategy for mitigating hallucinations and addressing outdated information in LLMs for fact-checking applications is Retrieval Augmented Generation (RAG). This method combines information retrieval with LLMs, enhancing their factual accuracy by integrating credible external sources, such as scientific databases and verified fact-checking sites (Leippold et al., 2024). Moreover, RAG systems can be fine-tuned by jointly training the retriever and the LLM, where the LLM provides supervisory signals for training the retriever (Izacard et al., 2023; Zeng and Gao, 2024). By combining retrieval and generation, RAG enhances the effectiveness of the fact-checking process against false information.

6 LLM Pipelines

In fact-checking, LLMs are commonly integrated into complex orchestrated pipelines, where the output of one step feeds into the next. Several studies have proposed such pipelines, combining multiple techniques from Section 5 to aid fact-checkers.

A common approach is to **decompose a claim into verifiable sub-claims and predict the veracity of each sub-claim**. Zhang and Gao (2023) employed LLMs to break down claims by generating a series of questions and answers. The LLM can then rely on its internal knowledge or retrieve external information to answer, assessing its confidence before making a prediction. Similarly, Wang and Shu (2023) instructed LLMs to define predicates and follow-up questions. After generating answers, the LLM utilize a context of predicted questions and their corresponding answers to assess each predicate's veracity, which is subsequently aggregated into a final prediction and explanation.

Alternatively, claims can be assessed without decomposition by **directly tasking LLMs to generate questions and answers for veracity prediction and to provide an overall assessment of the claim** (Chakraborty et al., 2023). The model then decides whether the claim is supported, refuted or lacks sufficient information. Another approach is to convert claims into binary yes-or-no questions, which are used to create interpretable logic clauses for debunking misinformation (Liu et al., 2024a).

Beyond natural language, Pan et al. (2023b) introduced a **programming-based LLM for creating verification programs**. This approach first employs a generative LLM to decompose claims into questions, after which a programming LLM generates a reasoning program for verification. The resulting code is executed with specific sub-task func-

tions to produce a final veracity prediction. This framework serves as a baseline for *fact-verification*.

Since LLMs tend to produce hallucinated and unfaithful explanations, generative LLMs can also be used to **enhance faithfulness** (Kim et al., 2024). This approach harnesses two LLM debaters, which iteratively search for errors and flaws in the explanation. Subsequently, all identified errors and proposed corrections from both debaters are used to adjust the final justification. This process is valuable in fact-checking, which requires convincing and plausible explanations and fact-check articles, as a human fact-checker normally would do.

7 Languages

We also examined the language coverage of each work. **Most papers focus on a single language**, primarily English (56×), with only one each for Chinese and Arabic. Only 11 papers addressed more than one language, typically pairing English with another. Notably, only three explored ten or more languages, with the largest experiments covering 114 languages (Setty, 2024).

Prompt language. Most researchers use approaches designed for single-language settings in multilingual contexts, typically providing **instructions in English**, while leaving the claims in their original languages (Du et al., 2022; Agrestia et al., 2022; Hyben et al., 2023; Cao et al., 2023; Huang and Sun, 2023; Pelrine et al., 2023; Li and Zhai, 2023). This allows for consistent prompts without adaption. Alternatively, **instructions can be in the claim's language**, which requires predefined instructions for each language and an accurate language detector.

Multilingual training. An alternative approach involves **fine-tuning LLMs on multilingual data** to create models capable of handling multiple languages (Quelle and Bovet, 2023). Another possible strategy is **cross-lingual transfer**, where models are trained on high-resource languages and then evaluated in other languages.

8 Future Directions and Challenges

Knowledge-Augmented Strategies. Techniques incorporating external knowledge into LLMs have shown promise in addressing complex tasks in the NLP domain, including fact-checking. Despite their potential, these techniques remain underexplored in fact-checking, with only five out of 69 surveyed papers employing external sources. Future research should focus on leveraging these techniques, e.g., RAG, to enhance the efficiency and accuracy of LLM-driven fact-checking.

Multilingual Fact-Checking. Given the global nature of false information, leveraging LLMs for fact-checking across languages is essential. Developing effective multilingual fact-checking techniques will enhance the ability of LLMs to detect and address misinformation in diverse linguistic contexts. One challenge includes the underexplored task of *previously fact-checked claims detection*, which also involves cross-lingual claim-matching – where input claims and fact-checks are in different languages. This presents a valuable opportunity for further research, highlighting the potential of LLMs in multilingual fact-checking.

Fact-Checking in Real Time. Current methods are mostly reactive, addressing claims after they gain traction on social media. Generative LLMs could enable real-time monitoring and analysis, providing continuous updates on emerging information. Integrating LLMs with real-time information will help to identify and flag false information as it arises, enhancing the timeliness and accuracy of false claim detection and mitigating misinformation more effectively.

Interactive Fact-Checking. Interactive fact-checking tools powered by generative LLMs represent a promising direction for users to verify claims through dynamic dialogues. These systems could facilitate deeper engagement with fact-checking, offering explanations and follow-up questions. Future work should explore the development of interactive fact-checking tools to empower fact-checkers in false information mitigation.

9 Conclusion

The rapid advancement of generative LLMs has sparked considerable interest in their potential applications within the fact-checking domain. This study presents a systematic review of 69 research papers, providing a detailed analysis of the various methods that incorporate generative LLMs into the information verification process. By examining the techniques and approaches explored in the existing literature, this survey offers a comprehensive overview of current methodologies and a foundation for future research efforts to enhance and explore new frontiers in LLM-assisted fact-checking.

Limitations

Focus on Generative LLMs. This survey paper focuses exclusively on approaches and techniques employed in fact-checking that leverage generative LLMs. While various types of language models exist, such as encoder-only models, recent research in NLP has increasingly centered on generative LLMs due to their rapid advancements and versatility across a wide range of tasks. Although other model architectures could contribute to fact-checking, this paper emphasizes generative LLMs due to their growing prominence and potential in this domain.

LLMs for Fact-Checking. Using LLMs in fact-checking intersects with several related domains, including hallucination, LLM factuality or truth discovery. These domains address the issue of false information from various perspectives. However, most research on hallucination and LLM factuality aims to evaluate the model's outputs and determine whether there are discrepancies between generated answers and real-world facts. In contrast, this survey focuses on approaches and techniques for verifying information rather than evaluating LLM outputs. Additionally, we include methods for improving the faithfulness of LLM-generated explanations to enhance the reliability of fact-checking.

Textual Information Only. False information extends beyond text and can involve various modalities, such as images, videos, and audio. However, this survey is limited to the problem of verifying textual information. We excluded studies focused on multimodal approaches, thereby narrowing our scope to fact-checking methods that involve text processing only.

References

S Agrestia, AS Hashemianb, and MJ Carmanc. 2022. PoliMi-FlatEarthers at CheckThat! 2022: GPT-3 applied to claim detection. *Working Notes of CLEF*.

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Firoj Alam, Alberto Barrón-Cedeño, Gullal S Cheema, Sherzod Hakimov, Maram Hasanain, Chengkai Li, Rubén Míguez, Hamdy Mubarak, Gautam Kishore Shahi, Wajdi Zaghouani, et al. 2023. Overview of the clef-2023 checkthat! lab task 1 on check-worthiness in multimodal and multigenre content. *Working Notes of CLEF*.

Saud Althabiti, Mohammad Ammar Alsalka, and Eric Atwell. 2023. Generative ai for explainable automated fact checking on the factex: A new benchmark dataset. In *Multidisciplinary International Symposium on Disinformation in Open Online Media*, pages 1–13. Springer.

Saud Althabiti, Mohammad Ammar Alsalka, and Eric Atwell. 2024. Ta'keed: The first generative fact-checking system for arabic claims. *arXiv preprint arXiv:2401.14067*.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A benchmark dataset of check-worthy factual claims. *arXiv preprint arXiv:2004.14425*.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Alberto Barrón-Cedeno, Tamer Elsayed, Preslav Nakov, Giovanni Da San Martino, Maram Hasanain, Reem Suwaileh, and Fatima Haouari. 2020. Checkthat! at clef 2020: Enabling the automatic identification and verification of claims in social media. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II 42*, pages 499–507. Springer.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2021. Automatic claim review for climate science via explanation generation. *arXiv preprint arXiv:2107.14740*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric

Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Mars Gokturk Buchholz. 2023. Assessing the Effectiveness of GPT-3 in Detecting False Political Statements: A Case Study on the LIAR Dataset. *arXiv preprint arXiv:2306.08190*.

Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou, and Songlin Hu. 2023. Are large language models good fact checkers: A preliminary study. *arXiv preprint arXiv:2311.17355*.

Recep Firat Cekinel and Pinar Karagoz. 2024. Explaining veracity predictions with evidence summarization: A multi-task model approach. *arXiv preprint arXiv:2402.06443*.

Mohna Chakraborty, Adithya Kulkarni, and Qi Li. 2023. An empirical study of using chatgpt for fact verification task. *arXiv preprint arXiv:2311.06592*.

Canyu Chen and Kai Shu. 2023a. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788*.

Canyu Chen and Kai Shu. 2023b. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3495–3516, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Eun Cheol Choi and Emilio Ferrara. 2023. Automated claim matching with large language models: Empowering fact-checkers in the fight against misinformation. *arXiv preprint arXiv:2310.09223*.

Eun Cheol Choi and Emilio Ferrara. 2024. Fact-gpt: Fact-checking augmentation via claim matching with llms. *arXiv preprint arXiv:2402.05904*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Charles LA Clarke, Saira Rizvi, Mark D Smucker, Maria Maistro, and Guido Zuccon. 2020. Overview of the trec 2020 health misinformation track. In *TREC*.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *arXiv preprint arXiv:2006.00885*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2021. Climate-fever: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.

SM Du, Sujatha Das Gollapalli, and See-Kiong Ng. 2022. Nus-ids at checkthat! 2022: identifying check-worthiness of tweets using checkthat5. *Working Notes of CLEF*.

Julian Eisenschlos, Bhuwan Dhingra, Jannis Bulian, Benjamin Börschinger, and Jordan Boyd-Graber. 2021. Fool me twice: Entailment from Wikipedia gamification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 352–365, Online. Association for Computational Linguistics.

Michael Evans, Dominik Soós, Ethan Landers, and Jian Wu. 2023. Msvec: A multidomain testing dataset for scientific claim verification. In *Proceedings of the Twenty-Fourth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, MobiHoc '23, page 504–509, New York, NY, USA. Association for Computing Machinery.

Marcos Fernández-Pichel, David E. Losada, and Juan C. Pichel. 2022. A multistage retrieval system for health-related misinformation detection. *Engineering Applications of Artificial Intelligence*, 115:105211.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed ELsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, and Heng Ji. 2022. NewsClaims: A new benchmark for claim detection from news with attribute knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6002–6018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2023. Ambifc: Fact-checking ambiguous claims with evidence. *arXiv preprint arXiv:2104.00640*.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2023. Language models hallucinate, but may excel at fact verification. *arXiv preprint arXiv:2310.14564*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022a. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022b. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 675–682, Online. Association for Computational Linguistics.

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. LESA: Linguistic encapsulation and semantic amalgamation based generalised claim detection from online content. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3178–3188, Online. Association for Computational Linguistics.

Sai Gurrapu, Lifu Huang, and Feras A. Batarseh. 2022. Exclaim: Explainable neural claim verification using rationalization. In *2022 IEEE 29th Annual Software Technology Conference (STC)*. IEEE.

Fatima Haouari, Maram Hasanain, Reem Suwaileh, and Tamer Elsayed. 2021. ArCOV19-rumors: Arabic COVID-19 Twitter dataset for misinformation detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 72–81, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Emma Hoes, Sacha Altay, and Juan Bermeo. 2023a. Leveraging chat-gpt for efficient fact-checking.(2023). *Available on: https://doi. org/10.31234/osf. io/qnjkf and*.

Emma Hoes, Sacha Altay, and Juan Bermeo. 2023b. Using chatgpt to fight misinformation: Chatgpt nails 72% of 12,000 verified claims.

Andrea Hrckova, Robert Moro, Ivan Srba, Jakub Simko, and Maria Bielikova. 2024. Autonomation, not automation: Activities and needs of fact-checkers as a basis for designing human-centered ai systems.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Yue Huang and Lichao Sun. 2023. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. *arXiv preprint arXiv:2310.05046*.

Martin Hyben, Sebastian Kula, Ivan Srba, Robert Moro, and Jakub Simko. 2023. Is it indeed bigger better? the comprehensive study of claim detection lms applied for disinformation tackling. *arXiv preprint arXiv:2311.06121*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Bohan Jiang, Zhen Tan, Ayushi Nirmal, and Huan Liu. 2023. Disinformation Detection: An Evolving Challenge in the Age of LLMs. *arXiv preprint arXiv:2309.15847*.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring listwise evidence reasoning with t5 for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 402–410, Online. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023a. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023b. Wice: Real-world entailment for claims in wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Wei-Yu Kao and An-Zi Yen. 2024. How we refute claims: Automatic fact-checking through flaw identification and explanation. *arXiv preprint arXiv:2401.15312*.

Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. WatClaimCheck: A new dataset for claim entailment and inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304, Dublin, Ireland. Association for Computational Linguistics.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *arXiv preprint arXiv:2402.07401*.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Nayeon Lee, Yejin Bang, Andrea Madotto, and Pascale Fung. 2020. Misinformation has high perplexity. *arXiv preprint arXiv:2006.04666*.

Markus Leippold, Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Juerg Luterbacher, and Christian Huggel. 2024. Automated fact-checking of climate change claims with large language models. *arXiv preprint arXiv:2401.12566*.

Miaoran Li, Baolin Peng, and Zhu Zhang. 2023a. Self-checker: Plug-and-play modules for fact-checking with large language models. *arXiv preprint arXiv:2305.14623*.

Qifei Li and Wangchunshu Zhou. 2020. Connecting the dots between fact verification and fake news detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1820–1825, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yifan Li and ChengXiang Zhai. 2023. An exploration of large language models for verification of news headlines. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 197–206.

Zizhong Li, Haopeng Zhang, and Jiawei Zhang. 2023b. A revisit of fake news dataset with augmented fact-checking by chatgpt. *arXiv preprint arXiv:2312.11870*.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024a. Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection. *arXiv preprint arXiv:2402.07776*.

Qiang Liu, Xiang Tao, Junfei Wu, Shu Wu, and Liang Wang. 2024b. Can large language models detect rumors on social media? *arXiv preprint arXiv:2402.03916*.

Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. 2015. Real-time rumor debunking on twitter. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1867–1870, New York, NY, USA. Association for Computing Machinery.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2024. An empirical study of catastrophic forgetting in large language models during continual fine-tuning.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. 2023. Ex-fever: A dataset for multi-hop explainable fact verification. *arXiv preprint arXiv:2310.09754*.

Huanhuan Ma, Weizhi Xu, Yifan Wei, Liuji Chen, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. 2024. Ex-fever: A dataset for multi-hop explainable fact verification. *arXiv preprint arXiv:2310.09754*.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3818–3824. AAAI Press.

Christopher Malon. 2021. Team papelo at FEVEROUS: Multi-hop evidence pursuit. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 40–49, Dominican Republic. Association for Computational Linguistics.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Shujaat Mirza, Bruno Coelho, Yuyuan Cui, Christina Pöpper, and Damon McCoy. 2024. Global-liar: Factuality of llms over time and geographic regions. *arXiv preprint arXiv:2401.17839*.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, et al. 2022. Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets. In *2022 Conference and Labs of the Evaluation Forum, CLEF 2022*, pages 368–392. CEUR Workshop Proceedings (CEUR-WS. org).

Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barr'on-Cedeno, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021a. Automated fact-checking for assisting human fact-checkers. *ArXiv*, abs/2103.07769.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas

Mandl, Mucahid Kutlu, and Yavuz Selim Kartal. 2021b. Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. *arXiv preprint arXiv:2109.12987*.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information; Knowledge Management*, CIKM '21. ACM.

Anna Neumann, Dorothea Kolossa, and Robert M Nickel. 2023. Deep learning-based claim matching with multiple negatives training. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 134–139, Online. Association for Computational Linguistics.

Wojciech Ostrowski, Arnav Arora, Pepa Atanasova, and Isabelle Augenstein. 2021. Multi-hop fact checking of political claims. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3892–3898. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Liangming Pan, Xinyuan Lu, Min-Yen Kan, and Preslav Nakov. 2023a. Qacheck: A demonstration system for question-guided multi-hop fact-checking. *arXiv preprint arXiv:2310.07609*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023b. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. *Fighting an Infodemic: COVID-19 Fake News Dataset*, page 21–29. Springer International Publishing.

Kellin Pelrine, Meilina Reksoprodjo, Caleb Gupta, Joel Christoph, and Reihaneh Rabbany. 2023. Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. *arXiv preprint arXiv:2305.14928*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smoleň, Martin Melišek, Ivan Vykopal, Jakub Simko, Juraj Podroužek, and Maria Bielikova. 2023a. Multilingual previously fact-checked claim retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16477–16500, Singapore. Association for Computational Linguistics.

Matúš Pikuliak, Ivan Srba, Robert Moro, Timo Hromadka, Timotej Smolen, Martin Melisek, Ivan Vykopal, Jakub Simko, Juraj Podrouzek, and Maria Bielikova. 2023b. Multilingual previously fact-checked claim retrieval. *arXiv preprint arXiv:2305.07991*.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2020. Scientific Claim Verification with VERT5ERINI. *arXiv preprint arXiv:2010.11930*.

Ronak Pradeep, Xueguang Ma, Rodrigo Nogueira, and Jimmy Lin. 2021. Vera: Prediction techniques for reducing harmful misinformation in consumer health search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2066–2070, New York, NY, USA. Association for Computing Machinery.

Nestor Prieto-Chavana, Julie Weeds, and David Weir. 2023. Automated query generation for evidence collection from web search engines. *arXiv preprint arXiv:2303.08652*.

Dorian Quelle and Alexandre Bovet. 2023. The perils & promises of fact-checking with large language models. *arXiv preprint arXiv:2310.13549*.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark. Association for Computational Linguistics.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the GermEval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12, Duesseldorf, Germany. Association for Computational Linguistics.

Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2021. Searching for scientific evidence in a pandemic: An overview of trec-covid. *arXiv preprint arXiv:2104.09632*.

Daniel Russo, Shane Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023a. Countering misinformation via emotional response generation. In *Pro-*

ceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 11476–11492, Singapore. Association for Computational Linguistics.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023b. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023c. Benchmarking the Generation of Fact Checking Explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

Soumya Sanyal, Tianyi Xiao, Jiacheng Liu, Wenya Wang, and Xiang Ren. 2024. Minds versus machines: Rethinking entailment verification with language models. *arXiv preprint arXiv:2402.03686*.

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marcin Sawiński, Krzysztof Węcel, Ewelina Paulina Księżniak, Milena Stróżyna, Włodzimierz Lewoniewski, Piotr Stolarski, and Witold Abramowicz. 2023. Openfact at checkthat! 2023: Head-to-head gpt vs. bert-a comparative study of transformers language models for the detection of check-worthy claims. In *CEUR Workshop Proceedings*, volume 3497.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *arXiv preprint arXiv:2305.13117*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *arXiv preprint arXiv:2103.08541*.

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3419–3425, Hong Kong, China. Association for Computational Linguistics.

Vinay Setty. 2024. Surprising efficacy of fine-tuned transformers for fact-checking over larger language models. *arXiv preprint arXiv:2402.12147*.

Shaden Shaar, Nikolay Babulkov, Giovanni Da San Martino, and Preslav Nakov. 2020. That is a known lie: Detecting previously fact-checked claims. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3607–3618, Online. Association for Computational Linguistics.

SD-H Michael Shliselberg and Shiri Dori-Hacohen. 2022. Riet lab at checkthat! 2022: improving decoder based re-ranking for claim matching. *Working Notes of CLEF*, pages 05–08.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. Fakenewsnet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Iknoor Singh, Carolina Scarton, and Kalina Bontcheva. 2023. Utdrm: unsupervised method for training debunked-narrative retrieval models. *EPJ Data Science*, 12(1):59.

Ivan Srba, Branislav Pecher, Matus Tomlein, Robert Moro, Elena Stefancova, Jakub Simko, and Maria Bielikova. 2022. Monant medical misinformation dataset: Mapping articles to fact-checked claims. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2949–2959, New York, NY, USA. Association for Computing Machinery.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32–43.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Lichao Sun, Yue Huang, Haoran Wang, Siyuan Wu, Qihui Zhang, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. 2024. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. 2023. From chaos to clarity: Claim normalization to empower fact-checking. *arXiv preprint arXiv:2310.14338*.

Xin Tan, Bowei Zou, and Ai Ti Aw. 2023. Evidence-based interpretable open-domain fact-checking with large language models. *arXiv preprint arXiv:2312.05834*.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Hoai Nam Tran and Udo Kruschwitz. 2022. ur-iw-hnt at checkthat! 2022: cross-lingual text summarization for fake news detection. *Working Notes of CLEF*.

Tyler Vergho, Jean-Francois Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Comparing gpt-4 and open-source language models in misinformation mitigation. *arXiv preprint arXiv:2401.06920*.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Ivan Vykopal, Matúš Pikuliak, Ivan Srba, Robert Moro, Dominik Macko, and Maria Bielikova. 2024. Disinformation capabilities of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14830–14847, Bangkok, Thailand. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity.

Gengyu Wang, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown. 2023b. Check-COVID: Fact-checking COVID-19 news claims with scientific evidence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14114–14127, Toronto, Canada. Association for Computational Linguistics.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. *arXiv preprint arXiv:2310.05253*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. Factuality of large language models in the year 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jiaying Wu and Bryan Hooi. 2023. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. *arXiv preprint arXiv:2310.10830*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Fengzhu Zeng and Wei Gao. 2023. Prompt to be consistent is better than self-consistent? few-shot and zero-shot fact verification with pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4555–4569, Toronto, Canada. Association for Computational Linguistics.

Fengzhu Zeng and Wei Gao. 2024. Justilm: Few-shot justification generation for explainable fact-checking of real-world claims. *arXiv preprint arXiv:2401.08026*.

Xia Zeng, Amani S Abumansour, and Arkaitz Zubiaga. 2021. Automated fact-checking: A survey. *Language and Linguistics Compass*, 15(10):e12438.

Xia Zeng and Arkaitz Zubiaga. 2024. Maple: Micro analysis of pairwise language evolution for few-shot claim verification. *arXiv preprint arXiv:2401.16282*.

Caiqi Zhang, Zhijiang Guo, and Andreas Vlachos. 2024a. Do we need language-specific fact-checking models? the case of chinese. *arXiv preprint arXiv:2401.15498*.

Hangwen Zhang, Qingyi Si, Peng Fu, Zheng Lin, and Weiping Wang. 2024b. Are large language models table-based fact-checkers? *arXiv preprint arXiv:2402.02549*.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. *arXiv preprint arXiv:2310.00305*.

A Methodology

To search for research papers, we employed mostly search on ACL Anthology[1], ArXiv and Google Scholar. Since the fact-checking domain is still evolving and there are multiple names for the same fact-checking tasks, we used task names from previous surveys on fact-checking. Based on fact-checking tasks, we defined a list of keywords for the search that are shown in Table 1. The search for articles was completed in February 2024. After filtering out the papers that are not directly related to fact-checking, are focused on fact-checking the LLM's output or do not use generative LLMs for fact-checking, we identified a list of 69 articles.

Keywords
fact-checking
check-worthy claim detection
claim detection
previously fact-checked claims detection
previously fact-checked claims retrieval
claim-matching
evidence retrieval
factuality prediction
claim verification
fact verification
fake-news detection
justification prediction
explanation generation
fact-checking dataset
disinformation detection
misinformation detection
rumour verification
rumour detection

Table 1: A list of keywords used when searching for research papers.

B Datasets & Languages

After the rise of social media, the necessity for the development of datasets aimed at verifying information began to emerge. The first existing datasets started to appear in early 2010 and their availability increased over time. Many of these datasets have been involved in experiments using LLMs in fact-checking. Several of them come from *Conference and Labs of the Evaluation Forum (CLEF)* and specifically *CheckThat! Lab* focused on various tasks related to information verification, political bias of news articles, and anti-social behaviour, such as hate speech or propaganda detection.

In addition, multiple datasets focused thematically on different topics, e.g. COVID-19, the climate crisis, or political claims. Several datasets were developed from data collected from fact-checking organizations, for example, MultiClaim (Pikuliak et al., 2023b) or LIAR (Wang, 2017). Furthermore, multiple authors engaged their own data collection resources, such as PolitiFact[2] or Google Fact Check Tools[3], which collects verified claims from the Internet based on the ClaimReview schema[4].

A key attribute for identifying false claims in cross-lingual settings is the existence of multilin-

[1] https://github.com/acl-org/acl-anthology/

[2] https://www.politifact.com/

[3] https://toolbox.google.com/factcheck/explorer

[4] https://schema.org/ClaimReview

gual datasets, comprising multiple languages or a combination of language-specific datasets. Previously, most datasets have focused on one or a few languages, specifically, English and Arabic are the most represented languages. Recently, there have been efforts to create multilingual datasets that include a broader range of languages, including e.g. MultiClaim (Pikuliak et al., 2023b) or X-Fact (Gupta and Srikumar, 2021). A list of all identified datasets employed for the LLM experiments is shown in Table 2.

C Fact-Checking Benchmarks

Several benchmarks have been developed to assess the performance of LLMs in fact-checking tasks. For example, Bang et al. (2023) focused on assessing ChatGPT's reasoning, hallucination, and interactivity capabilities across eight tasks, including misinformation detection. Another benchmark, TRUSTLLM, scrutinized the trustworthiness and safety features of 30 LLMs (Sun et al., 2024).

D Model Analysis

Various researchers have focused on different types of LLMs within the reviewed papers. In contrast, these could be divided into two main categories: **Encoder-Decoder** (Seq2Seq) and **Decoder-only**. Although the first Seq2Seq models already emerged in 2019, their first applications in fact-checking were seen in 2021, when the problem of automated fact-checking began to receive increasing attention. The most commonly used Seq2Seq models include the T5 model, while we also encounter the T0 (Sanh et al., 2022), Flan-T5 (Chung et al., 2022), or mT5 (Xue et al., 2021).

In contrast, the second category includes Decoder-only LLMs, which we further divided into **GPT-base**, **LLaMA-based** and **Other models**, similarly to Minaee et al. (2024).

1. **The GPT Family.** Generative Pre-trained Transformers (GPT) is a family of LLMs developed by OpenAI and includes mostly closed-source models, accessible by paid API. This category includes all GPT models (e.g. GPT-3, GPT-3.5, or GPT-4) or CODEX. In addition to these closed-source models, there are also several open-sourced models, such as the GPT-Neo model.

2. **The LLaMA Family.** The LLaMA family is another collection of LLMs based on the LLaMA or Llama2 foundation models developed by Meta. This category also includes other models derived from LLaMA models, such as Alpaca, Vicuna, or even Mistral.

3. **Other models.** The last category consists of models that do not fall under the GPT or LLaMA family. Within this category, various models have gained the attention of researchers. Examples include the multilingual BLOOM or the OPT, PaLM, and PaLM2 models. Along with these models, several authors also investigated Atlas, a retrieval-augmented language model.

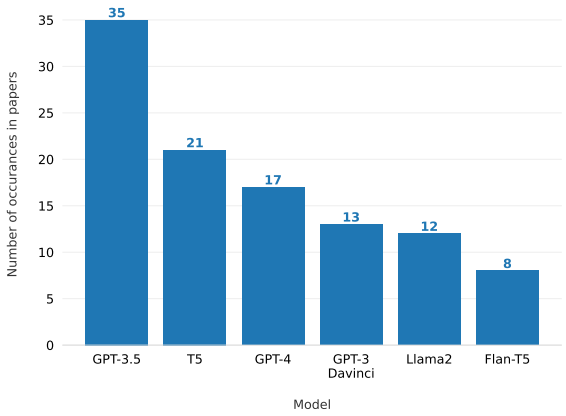

Figure 2: The occurrence frequency of the models in the reviewed papers. We only present models used in more than five papers.

Figure 2 depicts LLMs employed in more than five papers, accompanied by the respective count of papers each model utilized. A total of 33 various LLMs were employed in the collection of 69 papers.

Name	Claim detection	Previously fact-checked claims detection	Evidence retrieval	Fact verification	# Lang.	Size	Citation
ArCOV19-Rumors				✓	1	138	Haouari et al. (2021)
ArFactEx			✓	✓	1	100	Althabiti et al. (2024)
AveriTec				✓	1	5k	Schlichtkrull et al. (2023)
BoolQ-FV / AmbiFC				✓	1	11k	Glockner et al. (2023)
ChatGPT-FC				✓	1	22k	Li et al. (2023b)
Check-COVID			✓	✓	1	2k	Wang et al. (2023b)
CHEF				✓	1	10k	Hu et al. (2022)
ClaimBuster	✓				1	24k	Arslan et al. (2020)
ClaimDecomp	✓				1	1k	Chen et al. (2022)
CLAN	✓				1	6k	Sundriyal et al. (2023)
CLEF-2020	✓	✓	✓	✓	3	13K	Barrón-Cedeno et al. (2020)
CLEF-2021	✓	✓		✓	5	22k	Nakov et al. (2021b)
CLEF-2022	✓	✓		✓	7	36k	Nakov et al. (2022)
CLEF-2023	✓				3	63k	Alam et al. (2023)
CLIMATE-FEVER				✓	1	2k	Diggelmann et al. (2021)
Climate Feedback				✓	1	N/A	-
CoAID					1	5k	Cui and Lee (2020)
Constraint				✓	1	11k	Patwa et al. (2021)
COVID-19 Scientific				✓	1	142	Lee et al. (2020)
COVID-Fact			✓	✓	1	4k	Saakyan et al. (2021)
Data Common		✓		✓	X	N/A	-
e-FEVER				✓	1	68k	Stammbach and Ash (2020)
EnvClaims	✓				1	29k	Stammbach et al. (2023)
ExClaim				✓	1	4k	Gurrapu et al. (2022)
EX-FEVER				✓	1	61k	Ma et al. (2024)
FactEX				✓	1	12k	Althabiti et al. (2023)
FakeNewsNet				✓	1	23k	Shu et al. (2019)
FEVER			✓	✓	1	185k	Thorne et al. (2018)
FEVEROUS			✓	✓	1	87k	Aly et al. (2021)
FM2				✓	1	13k	Eisenschlos et al. (2021)
FullFact		✓		✓	1	N/A	-
GermEval 2021				✓	1	1k	Risch et al. (2021)
Global-LIAR				✓	1	600	Mirza et al. (2024)
HealthVer			✓	✓	1	14k	Sarrouti et al. (2021)
HoVer			✓	✓	1	26k	Jiang et al. (2020)
Labeled Unreliable News				✓	1	74k	Rashkin et al. (2017)
LESA-2021	✓				1	379k	Gupta et al. (2021)
LIAR				✓	1	13k	Wang (2017)
LIAR++				✓	1	6k	Russo et al. (2023b)
LIAR-NEW				✓	2	2k	Pelrine et al. (2023)
LLMFake					1	100	Chen and Shu (2023a)
Monant Medical Misinformation Dataset	✓	✓			1	51k	Srba et al. (2022)
MS-MARCO			✓		1	1011k	Bajaj et al. (2018)
MSVEC			✓	✓	1	200	Evans et al. (2023)
MultiClaim	✓	✓		✓	27/39	28k	Pikuliak et al. (2023a)
NewsClaims	✓				1	889	Gangi Reddy et al. (2022)
PolitiFact		✓		✓	1	N/A	-
PolitiHop				✓	1	500	Ostrowski et al. (2021)
PubHealth				✓	1	12k	Kotonya and Toni (2020)
RAWFC				✓	1	2k	Yang et al. (2022)
SciFact			✓	✓	1	1k	Wadden et al. (2020)
Snopes		✓		✓	1	N/A	-
Symmetric				✓	1	1k	Schuster et al. (2019)
TabFact				✓	1	18k	Chen et al. (2020)
TREC 2020 Health Misinfo.			✓		1	46	Clarke et al. (2020)
TREC-COVID			✓		1	171k	Roberts et al. (2021)
Twitter15				✓	1	2k	Liu et al. (2015)
Twitter16				✓	1	778	Ma et al. (2016)
VerMouth				✓	1	12k	Russo et al. (2023a)
VitaminC				✓	1	489k	Schuster et al. (2021)

WatClaimCheck	✓	1	34k	Khan et al. (2022)
Weibo21	✓	1	9k	Nan et al. (2021)
WiCE	✓	1	7k	Kamoi et al. (2023a)
X-Fact	✓	26	31k	Gupta and Srikumar (2021)

Table 2: A list of datasets utilized in experiments with LLMs for fact-checking.

Paper	Methods	Techniques	# Lang.
Gangi Reddy et al. (2022)	- classification - span detection	- fine-tuning - zero-shot - few-shot	1
Agrestia et al. (2022)	- binary classification - multilabel classification	- fine-tuning	4
Du et al. (2022)	- binary classification	- fine-tuning	6
Chen et al. (2022)	- question generation	- fine-tuning	1
Sawiński et al. (2023)	- binary classification	- fine-tuning - zero-shot - few-shot - few-shot CoT	1
Kamoi et al. (2023b)	- claim generation	- few-shot	1
Li et al. (2023a)	- claim generation	- few-shot	1
Hyben et al. (2023)	- binary classification	- fine-tuning - zero-shot - few-shot	10
Cao et al. (2023)	- binary classification	- zero-shot - few-shot - zero-shot CoT	2
Sundriyal et al. (2023)	- claim generation	- few-shot - zero-shot CoT	1
Setty (2024)	- binary classification	- zero-shot	114

Table 3: Work addressing the task of claim detection in the context of fact-checking. Each paper includes the utilized approaches, techniques and the number of languages analyzed.

Paper	Methods	Techniques	# Lang.
Shliselberg et al. (2022)	- ranking task	- fine-tuning	1
Singh et al. (2023)	- ranking task	- fine-tuning - zero-shot	1
Choi and Ferrara (2023)	- 3-class classification	- zero-shot - zero-shot CoT - few-shot CoT	1
Neumann et al. (2023)	- ranking task	- fine-tuning	1
Choi and Ferrara (2024)	- 3-class classification	- fine-tuning - zero-shot	1

Table 4: A list of papers addressing the task of detecting previously fact-checked claims. We include approaches, techniques and a number of languages for each paper.

Paper	Methods	Techniques	Knowledge base	# Lang.
Pradeep et al. (2020)	- ranking - rationale selection	- fine-tuning	- MS-MARCO	1
Pradeep et al. (2021)	- classification ranking	- fine-tuning	- given corpus	1
Sarrouti et al. (2021)	- classification ranking	- fine-tuning	- search engine	1
Jiang et al. (2021)	- classification ranking	- fine-tuning	- search engine	1
Fernández-Pichel et al. (2022)	- classification ranking	- fine-tuning	- Med-MARCO	1
Kamoi et al. (2023b)	- rationale selection	- fine-tuning	- Wikipedia	1
Prieto-Chavana et al. (2023)	- query generation	- fine-tuning - zero-shot - few-shot	- search engine	1
Schlichtkrull et al. (2023)	- question generation	- few-shot	- search engine	1
Wang et al. (2023b)	- rationale selection	- few-shot	- scientific papers	1
Evans et al. (2023)	- rationale selection	- zero-shot	- scientific papers	1
Tan et al. (2023)	- rationale selection	- few-shot	- search engine	1
Pan et al. (2023a)	- question generation	- few-shot	- Wikipedia - LLM knowledge	1
Setty (2024)	- question generation	- zero-shot	- search engine - fact-checks	114

Table 5: A list of relevant works focused on evidence retrieval. For each analyzed work we also present the knowledge base employed.

Paper	Methods	Techniques	Used Evidence	# Lang.
Pradeep et al. (2020)	- 3-class classification	- fine-tuning	- open-book	1
Petroni et al. (2021)	- 3-class classification	- fine-tuning	- open-book	1
Pradeep et al. (2021)	- 3-class classification	- fine-tuning	- open-book	
Sarrouti et al. (2021)	- 3-class classification	- fine-tuning	- open-book	1
Malon (2021)	- 3-class classification - next hop prediction	- fine-tuning	- open-book	1
Jiang et al. (2021)	- multi-class classification	- fine-tuning	- open-book	1
Tran and Kruschwitz (2022)	- 4-class classification	- fine-tuning	- open-book	2
Bhatia et al. (2021)	- binary classification - 3-class classification - explanation generation	- fine-tuning	- open-book - closed-book	1
Hoes et al. (2023a)	- binary classification - 6-class classification	- zero-shot	- closed-book	1
Kamoi et al. (2023b)	- binary classification	- fine-tuning - few-shot	- open-book	1
Schlichtkrull et al. (2023)	- 3-class classification - explanation generation	- few-shot	- open-book - closed-book	1
Buchholz (2023)	- 6-class classification	- fine-tuning - zero-shot	- closed-book	1
Wang et al. (2023b)	- 3-class classification	- few-shot	- open-book	1
Pan et al. (2023b)	- binary classification - text generation - explanation generation	- zero-shot - few-shot - zero-shot CoT - few-shot CoT - few-shot self-ask	- gold evidence - open-book - closed-book	1
Izacard et al. (2023)	- 3-class classification	- fine-tuning (RAG)	- open-book	1
Huang and Sun (2023)	- binary classification - 3-class classification - explanation generation	- zero-shot - zero-shot CoT - self-aware prompt	- closed-book	2
Jiang et al. (2023)	- binary classification - regression - explanation generation	- zero-shot	- closed-book	1
Chen and Shu (2023a)	- binary classification	- zero-shot - zero-shot CoT	- closed-book	1
Quelle and Bovet (2023)	- 4-class classification - 7-class classification	- ReAct	- open-book - closed-book	14
Ma et al. (2023)	- 3-class classification - explanation generation	- zero-shot - few-shot	- open-book	1
Guan et al. (2023)	- regression	- fine-tuning - zero-shot	- open-book - closed-book	1
Wu and Hooi (2023)	- binary classification	- zero-shot	- closed-book	1
Evans et al. (2023)	- 3-class classification - regression	- zero-shot	- open-book	1
Althabiti et al. (2023)	- explanation generation	- fine-tuning	- closed-book	1
Chakraborty et al. (2023)	- binary classification - 3-class classification - explanation generation	- zero-shot	- closed-book	1
Bang et al. (2023)	- explanation generation	- zero-shot	- closed-book	1
Tan et al. (2023)	- 3-class classification - explanation generation	- zero-shot	- open-book	1
Zeng and Gao (2023)	- 3-class classification	- fine-tuning - zero-shot - few-shot - self-consistency CoT	- open-book	1
Wang and Shu (2023)	- 3-class classification - text generation	- few-shot - CoT - self-ask - FOLK	- open-book	1
Pelrine et al. (2023)	- binary classification - regression - explanation generation	- fine-tuning - zero-shot - few-shot	- open-book - closed-book	3
Zhang and Gao (2023)	- 3-class classification - 6-class classification - text generation	- few-shot - few-shot CoT - search-augmented CoT - ReAct - HiSS	- open-book - closed-book	1
Li et al. (2023b)	- regression - explanation generation	- zero-shot	- closed-book	1

Hoes et al. (2023b)	- binary classification - 4-class classification	- zero-shot	- closed-book	1
Pan et al. (2023a)	- binary classification - explanation generation	- zero-shot - few-shot - few-shot CoT	- open-book - closed-book	1
Russo et al. (2023a)	- data generation	- zero-shot		1
Russo et al. (2023b)	- explanation generation	- fine-tuning	- open-book	1
Li and Zhai (2023)	- 6-class classification	- zero-shot	- closed-book	2
Vergho et al. (2024)	- binary classification - regression - explanation generation	- zero-shot	- closed-book	2
Zeng and Gao (2024)	- 3-class classification - explanation generation	- fine-tuning (RAG) - few-shot	- open-book	1
Leippold et al. (2024)	- binary classification - 5-class classification - 7-class classification - 12-class classification	- RAG	- open-book - closed-book	1
Kao and Yen (2024)	- explanation generation	- fine-tuning	- closed-book	1
Zhang et al. (2024a)	- 3-class classification	- few-shot	- gold evidence - open-book	1
Althabiti et al. (2024)	- 3-class classification - explanation generation	- zero-shot	- open-book	1
Zeng and Zubiaga (2024)	- 3-class classification	- fine-tuning - few-shot	- open-book	1
Mirza et al. (2024)	- binary classification - 3-class classification	- zero-shot	- closed-book	1
Sun et al. (2024)	- 3-class classification	- zero-shot	- open-book	1
Zhang et al. (2024b)	- binary classification	- fine-tuning - zero-shot - few-shot	- open-book	1
Cekinel and Karagoz (2024)	- explanation generation	- fine-tuning	- open-book	1
Liu et al. (2024b)	- binary classification - explanation generation	- zero-shot	- closed-book	1
Kim et al. (2024)	- binary classification - explanation generation	- MADR - zero-shot - CoT - self-refine	- open-book	1
Sanyal et al. (2024)	- binary classification	- zero-shot	- open-book	1
Setty (2024)	- regression - explanation generation	- zero-shot	- open-book	46
Liu et al. (2024a)	- binary classification - 6-class classification	- zero-shot - few-shot - zero-shot CoT - few-shot CoT - TELLER	- open-book - closed-book	1

Table 6: A comprehensive list of works for the fact verification task. Most of the surveyed papers address this task. In the table, we provide used evidence as open and closed-book, based on either the use of an external source or the knowledge embedded in the parameters.